Proceedings of the Doctoral Consortium at the
18th International Conference on Asia-Pacific Digital Libraries (ICADL 2016), and
Asia-Pacific Forum of Information Schools (APIS 2016)
December 6th, 2016

Concept, Process and Principle: Research on the Privacy Protection in Web Archiving

Heng HOU

School of Information Management,

Sun Yat-sen University,China

Email:houheng@mail2.sysu.edu.cn

Supervisor:Dr. Yongsheng CHEN

School of Information Management,

Sun Yat-sen University,China

Email: 13678990606@163.com

ABSTRACT

The purpose of this article is to clarify the privacy protection in web archiving, including the concept, the management process and the principle, which will provide a reference to protect the personal privacy in this big data era. It chooses the typical practices of archiving websites, such as the Preserving and Accessing Networked Documentary Resources of Australia, the UK Central Government Web Archive, making research on their archiving process, policies and responsibilities, and summarizes their common principle and emphasis. Then it identifies that there are two kinds of people that involve the archiving websites' privacy protection: the creators of the information and the users of the information. Also, there are three different types of privacy: personal information, communication privacy and accession privacy. It points out that there are four important process for the privacy protection during the whole websites archiving work: selection, preservation and accession. Also, the protection of the creators' individual information and the users' privacy are both significant. Finally, it summarizes four principles of the privacy protection in web archiving, pointing that web archiving is a complicated work which need cooperation and guidance.

INTRODUCTION

The World Wide Web began in 1993, which opened up online publishing as an easily available and ubiquitous (Phillips, 2005). With the rapidly development and expansion of

internet, nowadays almost everyone in the world involves in that and can set themselves up as a publisher. Information published on the Internet will be the primary resources, for our society and also for the future generation. It will be the vital evidence of our daily life, the official activities and also the heritage of our society. This not only means that the amount of online publishing is countless, but also impresses that there is huge significant information on the Internet. Therefore, it is necessary to preserve the websites. Web archiving has become the new issue in archival science.

Web archiving focuses on how to capture or "freeze"websites, which means that they document a moment in a website's lifetime, aiming at creating stable and enduring containers of the website's past(Ben-David, 2016).To a country, web archiving is to preserve a country's national activities and social heritage.

Compared with the traditional archives, different countries face different situations and have different issues when it comes to web archiving, so that the emphasis of privacy protection in different countries are various(International Internet Preservation Consortium,2016).However, there still have the basic common principle and that may make contribution to other countries when they start to make their own policies and strategies. This article chooses the typical practices of web archiving, including, Australia's Web Archive program PANDORA, the UK Central Government Web Archive, the International Internet Preservation Consortium and so on, and analyses the concepts, the manage processes and the policies of privacy protection, and the summarizes their management principles.

CONCEPT OF WEB ARCHIVING

What is web archiving?

As more and more public services are available on the Internet, the information on these websites record the activities of different kinds of organizations, which are the valuable evidences of administrative, fiscal and legal for these organizations. In addition, the information on the different kinds of websites reserve the vestiges of different organizations and individuals, which has the historical value for the society and next generation. To the national archiving work, capturing, preserving and accessing the websites is necessary .

Therefore, it is essential to archive the websites and clarify that web archiving is also the significant part of the national archiving work in these information era.

What is web archiving? One of the simple definitions is that a website or part of a website selected for preservation(UK Web Archive,2016). Web Archiving operates at the frontier of capturing and preserving our contemporary cultural and historical record(Sara,2016). It is used to mean the act of downloading from the Internet and storing on the national archive, and also the policies, guidelines, procedures and technical infrastructure that supports the archive(National Library of Australia,2013). More specifically, it is the process of acquiring data that has been published on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research. It is practiced internationally by memory institutions and private organizations to safeguard documentary heritage from the World Wide Web(International Internet Preservation Consortium,2016). Since it is a comprehensive work, many countries have built the program or specific website to carry out the web archiving work. The UK Web Archive contains websites that publish research, that reflect the diversity of lives, interests and activities throughout the UK, and demonstrate web innovation(The Government of Canada Web Archive,2016).In Australia, the PANDORA, Australia's Web Archive was established by the National Library in 1996 and is a collection of historic online publications relating to Australia and Australians. In Canada, since 2005, Library and Archives Canada (LAC) has collected federal and non-federal web resources in the context of its Web Archiving Program. The Government of Canada Web Archive (GCWA), provides access to archived federal websites. The GCWA contains the federal government information as originally published on the Web, the federal web pages that are no longer available to the public, and the indices to explore the collection by organization name and by URL(UK Web Archive,2016).So when we discuss the archiving websites, it is about the preservation of the worthy websites, and it is also content the policies, the guidance and the technologies and so on. Also, archiving websites is not a one-time work but a dynamic process, which can provide the evolution of a website over time.

Which kind of the website that should be archived?

The major types of national websites can be divided into three types: one is the websites of public services, which is the important part of the national records, because there are more and more organizations use the websites to carry out their daily work, the information on these websites is as important as the normal jobs off-line. Another type is the different professional websites, for example, the legal websites, the educational websites. There also another significant type, the social websites, which is the valuable record of a country and also the record of almost every individual. It may provide a rich historical resource that documents a time period in the way that fast-disappearing resources like local newspapers, city directories, and personal correspondence did in the past(Marshall &Shipman,2014）. However, a large amount of personal information on the social websites may makes it the hardest part of privacy protection in web archiving.

CONCEPT OF PRIVACY PROTECTION IN WEB ARCHIVING

What is privacy protection?

Privacy is defined as "the right of an individual to be secure from unauthorized intrusion and disclosure of information about oneself that is contained in documents, including the condition of not having undocumented personal knowledge about one possessed by others" (Parent, 1983). The protection of privacy is fundamentally about autonomy, power, and knowledge. The protection of personal privacy speaks to the ability of individuals to control what is known about themselves. individuals who can control what others know about them have a degree of personal autonomy. Privacy protection also helps to define the relationship between the citizen and the state and between the consumer and the corporate interest. Fundamentally, it attempts to balance what is best for the individual with what is best for the larger community the individual lives within(Stefanick, 2011）.

When it comes to the websites, the privacy protection is harder to defined, since the websites are opened for the public and almost everyone can create their own record on the website, and the information involves privacy is scattered and complex. Hence, it is more

difficult to clarify the privacy on the websites and also more difficult to make them to be appropriate used.

The different types of privacy information on the websites

The privacy in the websites is the protection of the confidentiality of their person data that is stored on the Internet and the rights of individuals not to be disturbed by unwanted information(Anderson & Weitz, 1992; Miller & Weckert, 2000). So there are two kinds of people that involve the web archiving's privacy protection: the creators of the information and the users of the information. Usually, individual that has contacted with the Internet palys these two roles, the creators and the users.

The privacy on the websites can be divided into three types: personal information, communication privacy and accession privacy. Firstly, personal information of the creators is that individuals put forward their publication on the Internet and any information about an identifiable individual. It is related to a person's private life or concerns, recorded in any form, including names, home address and telephone number, date of birth, social insurance number, age, marital and financial status, race, national or ethnic origin, religion, life styles, habits and behaviors (Van Riel& Jurriens,2001, Hung & Wong,2009). When it is necessary to capture the personal information on the websites, how to preserve and access is the important issue. Secondly, communication privacy is about the protection of data transmission and storage, which is very often done by encrypted files and communications (Elmer-Dewitt, 1994). Thirdly, accession privacy is when the users access the archives, how to protect their personal information of the accession needs to be clarify. Since that the data of accession is also one part of the archiving websites, during this period, how to deal with the personal information of the users is about the right to be left alone (Warren & Brandeis, 1890).

THE CRUCIAL PROCESSES OF PRIVACY PROTECTION IN WEB ARCHIVING

Like the normal archiving processes, web archiving also includes these basic manage processes, capture, preservation, and accession. However, specific to web archiving, it can be divided into more management processes. In this part, I will analyse the concrete processes and the critical job of privacy protection in each process of web archiving.

Selection

As with any other material that archives collect, web archiving intends to select the complement existing collections and serve different goals. Selection is the process that executed before capturing, which is aimed to clear the archiving scope. It is the first step and the fundamental process of web archiving, which defines what kind of information should be preserved and what kind of individual information should be protected. For the national web archiving, it often focuses on collection their national domains for heritage or as part of copyright deposit regimes and therefore perform broad, very large crawls (International Internet Preservation Consortium,2016).

In this process, the normal principle of different countries is that they do not want to archive all the information on the websites. On one hand, there are large volume of online resources, it is hard to archive all of that and have enough storage space. Like PANDORA, its selection policy is a pragmatic one of selecting at a level that will maximize access to important content, while taking in to account the workload involved. It depends on the web site. In the case of a very large web site, for example, the web site of a government department, it may select only a part of a site containing information about a particular initiative or program. Sometimes it selects only the publications from a very large site, for instance, an e-journal or newsletter, or scientific or technical reports (National Library of Australia,2013).On the other hand, it is not necessary to capture all the information on a website, since online publications is different from print publications. Publications are often mounted on the Internet without the quality filtering mechanism provided by editors and publishers. Consequently, standards are much more variable than they are in print. The social websites are the typical example. The UK Government Web Archive (UKGWA) points out that where material that forms part of the public record is published on sites or services outside of government ownership the aim is to retain this part of the record without infringing the rights of the site or service provider or other users of the service so capture and preservation is limited to material firmly in scope. For example, complex technical solutions have been developed in order to limit the risk of the capturing out of scope material when

archiving social media, and has put forward the detailed scope of that(The National Archives, 2014).PANDORA declares that the key areas that may involve the collection of personal information include : public access to items in the collection; engagement with the public via social media and online services; public events and activities; e-commerce activities; personnel management; physical and Information Technology security; and stakeholder liaison. And it will not seek information which it does not need(National Library of Australia,2016).

Last but not least, web archiving is not a one-time work, most of the websites are changing every second, harvesting the change is essential. The frequency of capture depends on the nature of the web site or publication, particularly its publication schedule, the value of the content, and how stable and long living the site is likely to be. Monographs need to be captured once only. There is at least one serial that is archived on a weekly basis. In the case of certain events, like Olympics, sites may be archived on a daily basis. Decisions are made on a case by case basis(National Library of Australia,2013).

Preservation

The purpose of the web archiving is providing long-term access to the heritage of society that published on the websites. For the web archiving, the collections are diverse in terms of types of collection materials, sources, and levels of control or influence over their creation. In addition, the digital collections are currently approaching petabyte scale, and are expected to continue growing. During the preservation, different information need for different preservation approaches, often at different scales, and possibly changing - preservation action cycles over time, using a changing suite of tools(The National Archives, 2013).For the privacy protection, it is to maintain the data should only be retained for the period necessary to perform the processing. Also, the data must be protected against unauthorized or unlawful processing, and from loss or damage. The level of security should be appropriate to the nature of the data and the potential harm that could arise from misuse (Brown,2006).

Accession

Web archives are born-digital collections that require special software tools for their use. Researchers can view archived web sites page-by-page or whole collections can be processed as data, revealing broad characteristics of the collections. The organizations affiliated with the IIPC are committed to ensure that their web archive collections are preserved and made accessible for future researchers, historians and the public (International Internet Preservation Consortium,2016).UK and Australia both emphasis that privacy legislation only applied to living persons. In UK, the Data Protection Act allows personal data to be preserved permanently for historical research purposes, if the data processing is not being used to support decisions about the data subject and is unlikely to cause substantial distress or damage (The National Archives, 2013).

Different access conditions may apply, depending on whether the content was freely accessible online or not. In UK, most, but not all, of the websites accessible through the UKGWA were created by Crown bodies and are Crown copyright. Most of the archived content of these websites and services is also Crown copyright. But Where websites have used third party (non-Crown) material the copyright status of this material should be clearly stated on the site, either attached to or embedded within the material itself or on the copyright page on said site. In such cases the third party content is not re-usable under the Open Government Licence and the onus for obtaining the consent of the copyright owner rests with the person or organisation who wishes to re-use it.

During the accession, the archiving websites will capture the information of usage and users at the same time. Normally, these data will only be used to improve the archiving work and the access services. The individual information should be preserved appropriately as well.

THE PRINCIPLES OF PRIVACY PROTECTION IN WEB ARCHIVING

The privacy protection of web archiving is not the same as the normal privacy protection, for the multiple sources of websites, a variety of formats and the diverse use restrictions, which is more complicated. Normally speaking, the websites are open to the public since they exist, so after archiving, the privacy issue of websites is more about the deep data and personal privacy, and not much about the content of the websites. According to the practices and policies in

different countries, the common principles privacy protection in web archiving can be summarized as following:

1.Cooperation is the important part of archiving websites. Because of the large volume of online resources to be archived and preserved and the high costs associated with this activity, web archiving is not a project that can be done by one organization, it is a project that need cooperation, so does the privacy protection work in websites archiving. PANDORA Archive has established a policy of cooperation with the State libraries and other cultural collecting agencies very early, and set up technical infrastructure, policy and procedures for archiving online publications and web sites and invited other collecting agencies to join it in building, which is stored centrally on a server at the National Library. It has entered into Memoranda of Understanding with participants. The National Library of Australia is working with other Commonwealth agencies which are responsible for the creation and management on online government information, GeoScience, the Australian Bureau of Statistics and the National Archives of Australia, to define responsibilities for preservation and long-term access to this data(National Library of Australia,2016).

2.Although there is privacy protection act in every country, it is still necessary to develop a specific guidance of that in archiving websites. The IIPC points that no one country is faced with the same challenges and legal situations, they different legal frameworks: some are awaiting legislation, others have legislation that covers web archiving, or other legal doctrines such as fair use (in some countries) that permit or mandate web archiving. Many web archiving organizations follow a permissions-based approach, in absence of legislation, or if the legal frameworks are unclear(International Internet Preservation Consortium,2016). The security of data,the data subject right,the copyright and transfer of information,the access purpose, etc, these problems exist in archiving websites need specific guidance. Under the national Privacy Protection Act, different countries should formulate their own regulations and guidance to protect the privacy on the websites, which are suitable to their own situation.

3.It should be clear that the different responsibilities of the organizations involve this job. In UK, government departments are responsible for maintaining their web presence in

accordance with all relevant legislation. Material published should comply with data protection, defamation and copyright law. The National Archives relies on the government organisations originally publishing the material on their sites to ensure that the content is Crown copyright and can be archived without infringing the copyright of non-government organisations or individuals. If appropriate permissions have not been granted by the copyright owners then The National Archives may be unable to archive a site (The National Archives, 2013).

It means that the responsibility of the organizations is to confirm their publications on their websites are legitimate, and when it comes to archiving, the responsibility of archives is to identify the legitimacy of the information in the websites, capture the information that conforms to the standard, guaranty them authenticity and long-term preservation, and provide appropriate accession. Before archiving, the privacy protection is the duty of the websites' owners, after archiving it becomes to the archives'.

4.It is essential to clarify the processes of websites archiving, which means that each action of this job should be clear at the beginning. The archiving of websites is more complex and need to harvest the same websites repeated, in this case, its management is different from the general digital archive. In the PANARO program, the National Library has developed a number of manuals and guides to assist its staff and other participants in undertaking the procedures involved. The national archives of UK has putted forward the Operational Selection Policy OSP 27: UK Central Government Web Estate, providing the guidance of each process and emphasis the privacy protection of that.

REFERENCES

Phillips, M. E. (2005). What should we preserve? The question for heritage libraries in a digital world. Library trends, 54(1), 57-71. Terms and Conditions of Use for the

Ben-David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. New Media & Society, 1461444816643790.

International Internet Preservation Consortium.(2016). About Archiving. Available at: http://www.netpreserve.org/web-archiving/about-archiving [accessed 15.09.16].

UK Web Archive. (2016). UK Web Archive statistics. Available at: http://www.webarchive.org.uk/ukwa/statistics [accessed 15.09.16].

Sara Day Thomson.(2016).Surveying the Domain: Three Days with the Web Archiving Team. Available at: http://britishlibrary.typepad.co.uk/webarchive/ [accessed 20.10.16].

National Library of Australia.(2013). FREQUENTLY ASKED QUESTIONS ABOUT PANDORA .Available at: http://pandora.nla.gov.au/panfaqs.html [accessed 15.09.16].

International Internet Preservation Consortium.(2016) .Why Archive the Web?. Available at: http://pandora.nla.gov.au/panfaqs.html [accessed 15.09.16].

The Government of Canada Web Archive.(2016). What is web archiving?Available at: http://www.bac-lac.gc.ca/eng/discover/archives-web-government/Pages/web-archives.aspx#b [accessed 18.09.16].

Marshall, C. C., & Shipman, F. M. (2014). An argument for archiving Facebook as a heterogeneous personal store. In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 11-20). IEEE Press.

Parent, W. A. (1983). Privacy, morality, and the law. Philosophy & Public Affairs, 269-288.

Stefanick, L. (2011). Controlling knowledge: Freedom of information and privacy protection in a networked world. Athabasca University Press.

Anderson, E., & Weitz, B. (1992). The use of pledges to build and sustain commitment in distribution channels. Journal of marketing research, 18-34.

Miller, S., & Weckert, J. (2000). Privacy, the Workplace and the Internet. Journal of Business Ethics, 28(3), 255-265.

Van Riel, A. C., Liljander, V., & Jurriens, P. (2001). Exploring consumer evaluations of e-services: a portal site. International Journal of Service Industry Management, 12(4), 359-377.

Hung, H., & Wong, Y. H. (2009). Information transparency and digital privacy protection: are they mutually exclusive in the provision of e-services?. Journal of Services Marketing, 23(3), 154-164.

Elmer-Dewitt, P. (1994). Nabbing the pirates of cyberspace. Time, 143(24), 62-63.

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.

National Library of Australia.(2016).POLICY AND PRACTICE STATEMENT. Available at: http://pandora.nla.gov.au/policy_practice.html [accessed 15.09.16].

The National Archives. (2014). OPERATIONAL SELECTION POLICY OSP27. Available at:http://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf [accessed 19.09.16].

National Library of Australia.(2016). PRIVACY POLICY.Available at: http://www.nla.gov.au/policy-and-planning/privacy-policy [accessed 15.09.16].

National Library of Australia.(2013). DIGITAL PRESERVATION POLICY 4TH EDITION (2013).Available at: http://www.nla.gov.au/policy-and-planning/digital-preservation-policy [accessed 25.09.16].

Brown, A. (2006). Archiving websites: a practical guide for information management professionals. Facet. 153

The National Archives. (2016). Information on web archiving. Available at: http://www.nationalarchives.gov.uk/webarchive/information.htm [accessed 19.09.16].

International Internet Preservation Consortium.(2016) .LEGAL ISSUES. Available at: http://netpreserve.org/web-archiving/legal-issues [accessed 15.09.16].